Q&A
20 October 2021
Explainable Machine Learning for Trustworthy AI
Dr Fosca Giannotti

Q1: How do you contextualise explainable AI for the case of mobility data especially when it comes to transport and urban planning? How can we evaluate the trust of policy-makers towards AI and ML techniques which they use for decision making?

OK, so the first for mobility data, that's one very, very interesting question. These data habilitate a variety of intelligent services aimed at supporting different decision makers: from the citizen that wants to know his personal best trip to the urban planner that needs to take decision on transportation policies. In both cases, is it possible to empower such users returning knowledge that often is the result of complex combination of data driven and model driven processes. To make such empowerment effective explanation is a requirement, so there is the need to feed the visual analytics interfaces with explanation of the recommendations coming by deep models. Current effort on our lab is extending our results on explanator for deep models for time series also to mobility data: local explanator that provide explanation in the form of exemplars and counter-exemplar.
So, how can we evaluate the trust of policymakers towards AI Machine learning techniques which they used for decision making? I think that we must change our attitude before deploying any AI systems, which kind of validation, which kind of trials we must do. We must be capable to validated the AI system also with respect to the kind of decision process, in particular we also need to measure the impact that explanations is capable to achieve (doing trials with and without explanations). This require new methodology and an important line of research that needs to involve other disciplines such as psychologist and sociologists - there are theories there that we can try to put in place with our techniques.

Q2: If a linear model can explain a deep neural network, would it have been better and equivalent to use the linear model in place of the deep neural network?

This is a very smart question so if you have a linear model, you must stay with the linear model. This is something that I'm very much convinced that if you can learn a transparent model by scratch you have to stay with that. Local explainers, are a good solution when globally, you are not capable of building a good surrogate (transparent) model. There is interesting research, which is working on the long term goal of having good transparent models, possibly integrating symbolic and sub symbolic reasoning, but still very far away. If the black box is very efficient because there are many features, so far is very difficult to have a transparent model built from scratch, that is equivalent to the black box.

Q3: Models like decision tree are unstable (the set of rules can change substantially upon minor changes in inputs) and have issues working with correlated inputs (select only one of them ignoring others). So, one may expect that there will be a lot of possible rule sets explaining the underlying black box model with the comparable accuracy. How do you solve these issues when using these models as surrogate ones to explain more complex models?

The set of rules can change substantially upon minor changes in the inputs.
Stability, fidelity, and faithfulness are important property for a local explainator. The way an explainator reconstructs the behaviour of the black box, includes some random step that may cause dramatic difference in generating explainations for multiple requests of same or similar instance. To avoid this, the design and implementation of the explainator itself needs to be very robust and stable with respect this specific issue.

Q4: As you have shown there are many different explanation models. How can we validate explanations from the models more rigorously and how do we know we can trust them with explaining new examples?

So, this is a very important line of research - validation. The validation part implies inventing, I say this word, inventing new methods for validating. Of course, I don't want to be too negative. So, there are methods to validate the explanation terms of quality in several metrics and also the accuracy with respect the black box model. So, in this sense, the methods are quite well formulated. It's much more difficult to evaluate the quality of the overall decision.